title

Tziporah Horowitz

Johns Hopkins University

Abstract

Your abstract here.

Introduction

Methods

There are several ways to measure the association between a risk factor and the binary outcome of contracting a disease. The following sections discuss four approaches in determining risk factors in the context of patients contracting heart disease. Sections 1–3 refer to figure 1 for simplification.

	Diseased	Healthy
Exposed	D_E	H_E
Unexposed	D_U	H_U

Figure 1. Contingency Matrix

For this analysis, the *Heart Disease Dataset* was collected from kaggle.com (Lapp, 2019). The dataset includes data that was compiled from four databases in 1988 and consists of 14 columns: 13 predictors and 1 target. The predictors include 5 continuous variables: age, resting blood pressure, serum cholestoral (in mg/dl), maximum heart rate achieved, and ST depression induced by exercise relative to rest (oldpeak); and 8 categorical variables: sex, chest pain type, fasting blood sugar > 120 mg/dl (true or false), resting electrocardiographic results, exercise induced angina (yes or no), the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, and thal (normal, fixed defect, reversible defect). Figure 2 shows the *skimpy* summary of all 14 variables and figure 3 shows the distributions of predictor variables when compared to the target, a binary indicator for the patient having heart disease.

1. Risk Difference

Often considered the simplest approach for measuring associated risk, *risk difference* or *absolute risk difference* (ARD) is the difference in the outcome rates between patients with the risk factor and patients without the risk factor (Telke & Eberly, 2011). Using the matrix in 1, risk difference can be defined mathematically as:

$$\text{ARD} = \frac{D_E}{D_E + H_E} - \frac{D_U}{D_U + H_U}$$

While the risk difference is easy to compute, its interpretation is often misleading and can only explain the associated risk between a single factor and the target.

2. Relative Risk

Similar to risk difference, *relative risk* compares the outcome rates between patients with the risk factor and patients without the risk factor. However, relative risk is computed as a ratio (RR) rather than a difference (Telke & Eberly, 2011). The risk ratio is defined as:

$$\mathrm{RR} = \frac{D_E/(D_E + H_E)}{D_U/(D_U + H_U)}$$

Relative risk is a useful statistic because it quantifies the probability of a patient with exposure contracting the disease relative to a patient without exposure. Risk ratios that are close to 1 indicate that the risk of contracting the disease for an exposed patient is the same as the risk for an unexposed patient. In contrast, risk ratios that are far from 1 indicate that there is an association between the variables. This allows one to create a confidence interval using the hypothesis test,

$$H_0 : RR = 1$$

$$H_1 : RR \neq 1$$

The risk ratio is considered a valid measure of relative risk in studies in which the sampling is dependent on the exposure of interest such as, randomized controlled trials or cohort and cross-sectional studies (Gallis & Turner, 2019). Like risk difference, relative risk can only explain the associated risk between a single factor and the target.

3. Odds Ratio

Often confused with risk ratio, *odds ratio* compares the statistical odds of the outcome in the exposed group to that of the outcome of the unexposed group. It is defined mathematically as:

$$\text{OR} = \frac{D_E/H_E}{D_U/H_U}$$

Like the risk ratio, odds ratios that are close to 1 indicate no association between exposure and contracting the disease, and odds ratios that are far from 1 indicate that there is an association between the variables. One can also create a confidence interval for the odds ration using a similar hypothesis test to that of the risk ratio, such that

$$H_0 : OR = 1$$

$$H_1 : OR \neq 1$$

While the odds ratio is typically considered the "only valid measure of relative association in traditional case-control studies" (Gallis & Turner, 2019), it is frequently misinterpreted as the risk ratio. However, in cases where the risk factor is relatively small ($< 10\%$), the odds ratio approximates the risk ratio:

$$\begin{matrix} \lim_{D_E \to 0} D_E + H_E = H_E \\ \lim_{D_U \to 0} D_U + H_U = H_U \end{matrix} \implies \frac{D_E/(D_E + H_E)}{D_U/(D_U + H_U)} \approx \frac{D_E/H_E}{D_U/H_U}$$

The odds ratio can be applied in multi-parameter settings when computed in a logistical regression analysis, due to its inherent calculation of the logit (or log-odds) function. To obtain the odds ratios of a logistic regression model, one simply has to exponentiate the coefficients.

4. Marginal Effects

Analysis

This is a Subsection

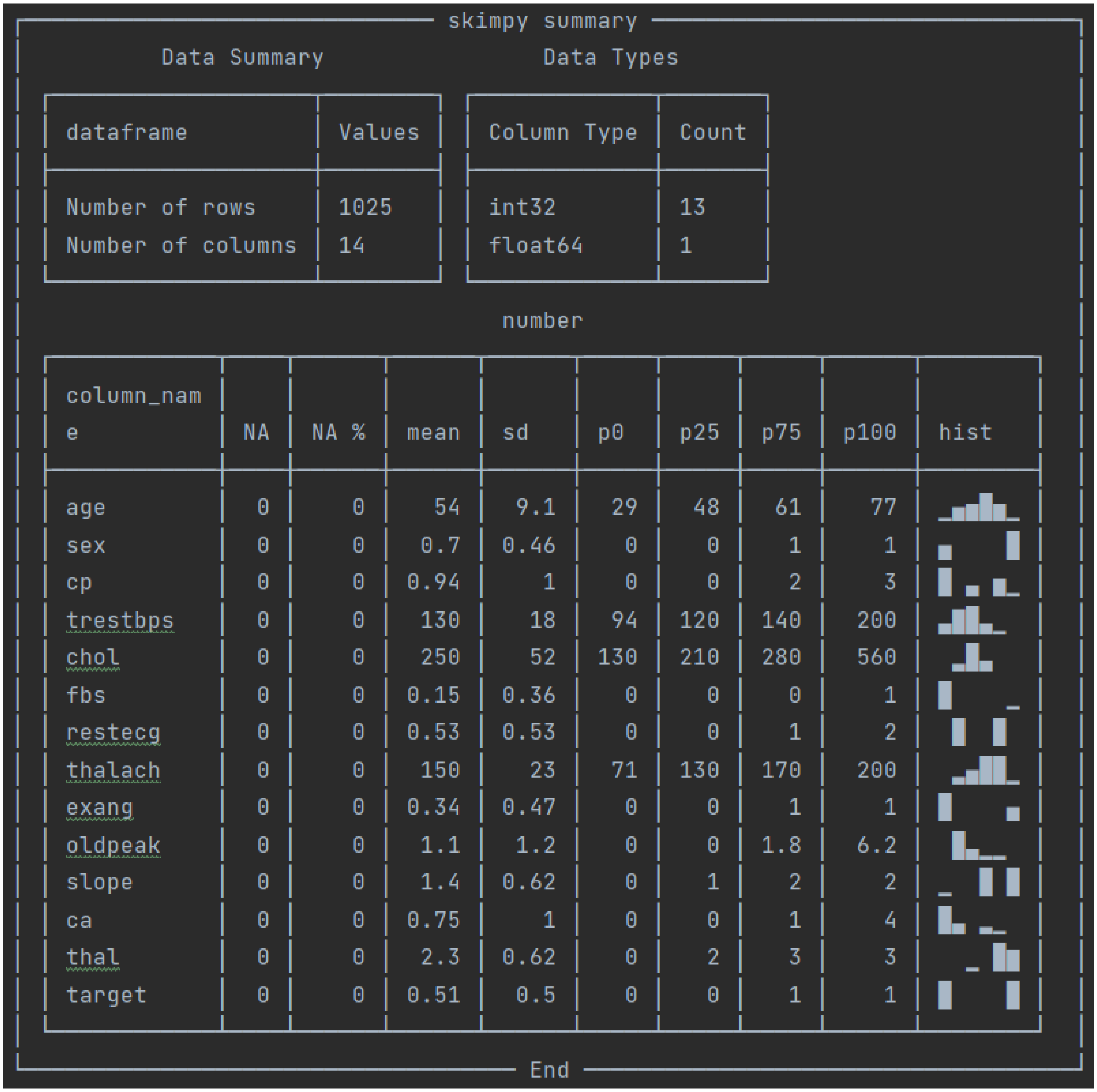
Results

References

Gallis, J. A., & Turner, E. L. (2019, November). Relative measures of association for binary outcomes: Challenges and recommendations for the global health researcher. *Ann Glob Health*, *85*(1), 137. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6873895/`

Lapp, D. (2019, Jun). *Heart disease dataset.* Retrieved from `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`

Telke, S. E., & Eberly, L. E. (2011, September). Statistical hypothesis testing: Associating patient characteristics with a prevalent or incident condition—relative risk, odds ratio, and logistic regression. *J. Wound Ostomy Continence Nurs.*, *38*(5), 496–500. Retrieved from `https://oce-ovid-com.proxy1.library.jhu.edu/article/00152192-201109000-00006/HTML`



skimpy summary

Data Summary

dataframe	Values
Number of rows	1025
Number of columns	14

Data Types

Column Type	Count
int32	13
float64	1

number

column_name	NA	NA %	mean	sd	p0	p25	p75	p100	hist
age	0	0	54	9.1	29	48	61	77	
sex	0	0	0.7	0.46	0	0	1	1	
cp	0	0	0.94	1	0	0	2	3	
trestbps	0	0	130	18	94	120	140	200	
chol	0	0	250	52	130	210	280	560	
fbs	0	0	0.15	0.36	0	0	0	1	
restecg	0	0	0.53	0.53	0	0	1	2	
thalach	0	0	150	23	71	130	170	200	
exang	0	0	0.34	0.47	0	0	1	1	
oldpeak	0	0	1.1	1.2	0	0	1.8	6.2	
slope	0	0	1.4	0.62	0	1	2	2	
ca	0	0	0.75	1	0	0	1	4	
thal	0	0	2.3	0.62	0	2	3	3	
target	0	0	0.51	0.5	0	0	1	1	

End

Figure 2. Summary of Variables

Figure 3. Distributions of Feature Variables